INDEPENDENT COMPONENT ANALYSIS FOR UNDERSTANDING MULTIMEDIA CONTENT

Thomas Kolenda, Lars Kai Hansen, Jan Larsen and Ole Winther
Informatics and Mathematical Modeling, Building 321
Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark
Phone: +45 4525 3920,3923,3889,3895
Fax: +45 4587 2599
E-mail: thko,lkh,jl,owi@imm.dtu.dk
Web: eivind.imm.dtu.dk

Abstract. This paper focuses on using independent component analysis of combined text and image data from web pages. This has potential for search and retrieval applications in order to retrieve more meaningful and context dependent content. It is demonstrated that using ICA on combined text and image features provides a synergistic effect, i.e., the retrieval classification rates increase if based on multimedia components relative to single media analysis. For this purpose a simple probabilistic supervised classifier which works from unsupervised ICA features is invoked. In addition, we demonstrate the use of the suggested framework for automatic annotation of descriptive key words to images.

INTRODUCTION

Understanding the structure of multimedia data is increasingly important for retrieval, indexing and search. With the advent of the advanced MPEG standards [25] content and context sensitive tools will become indispensable components of the webminers multimedia toolbox.

Content based image retrieval is a highly challenging aspect of multimedia analysis [9]. The task is hard because of the limited understanding of the relations between basic image features and abstract content descriptions. It is simply complicated to describe content in terms of intensity, edges, and texture. Therefore most current image retrieval systems, say on search engines like Google and FAST Multimedia Search, are based on analysis of an image and adjacent text on web page of the image.

Among the first commercial content based image retrieval systems worth mentioning are IBM's QBIC system [11], the VIR Image Engine from Virage, Inc. [12], and Visual RetrievalWare product by Excalibur Technologies [10]. These systems as well as the research prototypes mentioned in the reviews

[8, 9] aim at using primitive image features for retrieval. However, the most widely used image searches are primarily based on image associated keywords and adjacent text. If we want to perform more advanced searches it is required to invoke context sensitive text based approaches, i.e., invoke statistical tools like the vector space approach known as *latent semantic indexing* (LSI), see e.g. [5, 6].

We have argued earlier that independent component analysis (ICA) can be a valuable means for unsupervised structuring of multi media signals. ICA of text is also based on vector space representations, but do not search for orthogonal basis vectors as LSI and is not based on assumed multivariate normal statistics. In particular, we have shown that the independent components of text databases have intuitive meaning beyond that found by LSI, and similarly that independent components of image sets corresponds to intuitively meaningful groupings [13, 18, 19].

With this in mind we now explore independent component analysis of combined text and image data. We follow the approach taken by the search engines and use adjacency to associate text and images. If text and imagery are to mutually support each other, it is important that the independent components of the combined data do not dissociate. Indeed we will demonstrate that there is a synergistic effect, and that retrieval classification rates increase if based on multimedia components relative to single media analysis.

MODELING FRAMEWORK

Consider a collection of web pages consisting of images and adjacent text from which we want to perform unsupervised modeling, i.e., clustering into meaningful groups and possibly also supervised classification into labeled classes. Let $\boldsymbol{x} = [\boldsymbol{x}_I; \boldsymbol{x}_T]$ be the column vector of image (I) and text (T) features. Unsupervised ICA modeling, in principle, models the probability density $p(\boldsymbol{x})$ with the aim of identifying clusters in feature space. It has previously been shown that the projection onto an independent component subspace provides a meaningful clustering of features [19, 13]. The objective of supervised modeling is the conditional class-feature probability, $p(y|\boldsymbol{x})$, where $y = \{1, 2, \cdots, C\}$ is the class label. We will show that a simple probabilistic classifier can be combined with unsupervised ICA.

FEATURE EXTRACTION

Text Features

The so-called bag-of-words approach is used to represent the text. This approach is mainly motivated by its simplicity and its proven utility, see e.g., [6, 13, 18, 19, 26, 29], although more advanced statistical natural language processing techniques can be employed [24]. In text separation the data is

presented in the form of terms[1]. Each document, i.e., collection of terms adjacent to an image, is represented by a vector: the histogram of term occurrence, as purposed in the vector space model (VSM) [29]. The term vector is usually filtered by removing low and high frequency terms. Low frequency terms do not carry meaningful discriminative information. Similarly high frequency terms (also denoted stop-words) such as the, and or of are common to all documents. In this paper the stop-words were manually constructed to form a list of 585 words. Moreover, stemming is performed by merging words with different endings ed, ing or s. The collection of all document histograms provides the *term-document* matrix $\boldsymbol{X}_T = [\boldsymbol{x}_T(1), \cdots, \boldsymbol{x}_T(N)]$, where N is the number of documents.

Image Features

The intention is to employ VSM on image features, and previous work [5, 27, 30] indicate that the VSM in combination with latent semantic indexing (LSI) is useful. Thus we seek to construct a *feature-image* matrix $\boldsymbol{X}_I = [\boldsymbol{x}_I(1), \cdots, \boldsymbol{x}_I(N)]$. We suggested to use lowest level image features of the ISO/IEC MPEG-7 standard [25], which aims at setting the standards for multimedia content representation, understanding and encoding. The low level image features are color and texture which are implemented using HSV[2] encoding [30] and Gabor filters [23], respectively. In order to enhance sensitivity to the overall shape, e.g. background, each image is divided into 4×4 patches from which color and texture features are computed. Initial image subdivision experiments indicated that it is crucial for the overall performance.

Texture. By definition a texture is a spatially extended pattern build by repeating similar units called texels. Texture segmentation involves subdividing an image into differently textured regions. We use a Gabor filter bank where each filter output captures a specific texture frequency and direction. Basically Gabor filters have texture detecting capabilities [28, 15] which are motivated by the function of human primary visual cortex V1 and demonstrated to be independent components of natural scenes [3]. The Gabor filter impulse response is a Gaussian modulated by complex sinusoids [7],

$$h(n,m) = \frac{1}{2\pi\sigma^2}\, e^{-\frac{(n^2+m^2)}{2\sigma^2}}\, e^{j2\pi(Un+Vm)}, \tag{1}$$

where n, m are pixel indices. The filters in the bank are parameterized by the center frequency $f = \sqrt{U^2+V^2}$ which captures the repetition of the texels in the direction of the angular parameter $\theta = \tan^{-1}(V/U)$, and finally σ is the width parameter. In Figure 1, the Gabor filter impulse responses are shown. The filtered image patch is given as the 2D convolution $I_f(n,m) = I(n,m) * h(n,m)$, where I is the image patch. The texture features are then computed as the the total energy of the filtered outputs [16] $x_{IT}(f,\theta,\sigma) =$

[1] A term is one word or a small set of words that present a context.

[2] Hue, saturation and value.

$|\mathcal{P}|^{-1} \sum_{(n,m)\in\mathcal{P}} |I_f(n,m)|^2$, where $\mathcal{P}$ is the image patch with $|\mathcal{P}|$ pixels. The filter bank parameters defined as in [30] and are experimentally shown to be feasible.

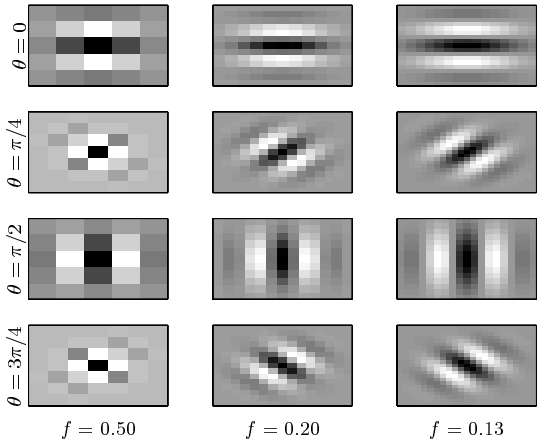


Figure 1: Gabor filters used to extract the texture features. Combining four directions $\theta = [0, \pi/4, \pi/2, 3\pi/4]$ and three frequencies $f = [0.50, 0.20, 0.13]$ gives a totoal of 12 filters in the bank. The width of the filters are $\sigma = 2$.

For each of the 16 images patches, the 12 energy texture features x_{texture} are computed and normalized to sum to one. This gives a total of $16 \cdot 12 = 192$ texture features. Finally the length of the 192 element texture feature vector is normalized to one.

Color. As in [30] we use the HSV (hue-saturation-value) color representation as color features. The HSV color space is believed to better linked to human color perception than e.g. standard RGB. The hue (H) can be interpreted as the dominant wavelength, S specify the saturation level, where zero corresponds to gray tone image. Finally, the value (V) specifies the lightness-darkness. Each color component is quantized into 16 bins, and each image patch is represented by 3 normalized color histograms. This gives 48 features for each of the 16 patches. In total, the color feature vector $\boldsymbol{x}_{IC}$ has $48 \cdot 16 = 768$ dimensions and is normalized to unit length.

INDEPENDENT COMPONENT ANALYSIS

The generative linear ICA model for the P-dimensional feature vector $\boldsymbol{x} = [\boldsymbol{x}_T; \boldsymbol{x}_{IT}; \boldsymbol{x}_{IC}]$ is given by:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}, \tag{2}$$

where $\boldsymbol{A}$ is the $P \times K$ mixing matrix, and $\boldsymbol{s} = [s_1, \cdots, s_k, \cdots, s_K]^\top$, $k = 1, 2, \cdots, K$ are $K \leq P$ independent sources. The literature suggest many

approaches for estimating the mixing matrix and the sources, such as: maximum likelihood optimization [2, 21], optimization of contrast functions from higher-order cumulants [4], kernel methods [1] and Bayesian learning [14, 20]. Due to its robustness and simplicity we will use the Infomax algorithm [2]. As suggested in [18, 19] latent semantic indexing (LSI) through Principal Component Analysis is suitable for projecting onto subspace. That is, the model is $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{\Phi}\boldsymbol{s}$, where $\boldsymbol{U}$ is the $P \times K$ matrix of K largest eigenvectors of the covariance of $\boldsymbol{x}$, and $\boldsymbol{\Phi}$ is the $K \times K$ mixing matrix. ICA is thus performed in the subspace $\widetilde{\boldsymbol{x}} = \boldsymbol{U}^\top \boldsymbol{x}$. The ICA model is estimated from a training set $\boldsymbol{X} = \boldsymbol{x}(1), \cdots, \boldsymbol{x}(N)$ of N related images/text data samples[3] to yield estimates $\widehat{\boldsymbol{U}}$, $\widehat{\boldsymbol{\Phi}}$.

The major advantage of combining ICA with LSI is that the sources are better aligned with meaningful content, which has been demonstrated for text documents in [19]. The different source components provide a meaningful segmentation of the feature space and mainly one source is active for a specific feature vector. That is, we can compute an estimated component conditional probability by softmax normalization,

$$\widehat{p}(k|\boldsymbol{x}) = \frac{\exp(\widehat{s}_k)}{\sum_{k=1}^{K} \exp(\widehat{s}_k)}, \;\; \widehat{\boldsymbol{s}} = [\widehat{s}_1, \cdots \widehat{s}_K]^\top = \widehat{\boldsymbol{\Phi}}^{-1} \widehat{\boldsymbol{U}}^\top \boldsymbol{x}. \tag{3}$$

Component Interpretation

In order to interpret the individual components, the K'th column of $\widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Phi}}$ will constitute text and image features associated with the K'th component/segment. Since the textual features are term-histograms we can further display high occurrence terms – keywords – which in the experimental section are demonstrated to yield meaningful interpretation of the components. In detail, we rank the terms according to probability and terms which above a certain threshold are reported as keywords. Similarly, high values of image features associated with a component provide a compact texture and color interpretation.

Probabilistic ICA Classification

Suppose that labels have been annotated to the data samples, i.e., we have a data set $\{\boldsymbol{x}(n), y(n)\}_{n=1}^{N}$ where $y(n) \in [1; C]$ are class labels. A simple probabilistic ICA classifier is then obtained as:

$$p(y|\boldsymbol{x}) = \sum_{k=1}^{K} p(y|k)p(k|\boldsymbol{x}), \tag{4}$$

where $p(k|\boldsymbol{x})$ is the conditional component probability estimated using ICA as given in Eq. (3). Provided that the independent components have been estimated, the conditional class-component probabilities, $p(y|k)$ are easily

[3] A pre-normalization, $\|\boldsymbol{x}\|_2 = 1$ is performed.

estimated from data as the frequency of occurrence for specific component-class combination $k \in [1; K]$, $y \in [1; C]$, as shown by

$$\widehat{p}(y|k) = \frac{1}{N} \sum_{n=1}^{N} \delta(\, y - y(n)\,) \cdot \delta(\, k - \arg\max_{k'} \widehat{p}(k'|\boldsymbol{x}(n))\,), \tag{5}$$

where $\delta(a) = 1$ if $a = 0$, and 1 otherwise. The stagewise training of the probabilistic classifier - which might be viewed as a mixture model - is suboptimal. All parameters in Eq. (4) should be estimated simultaneous, e.g., using a likelihood principle, however, the simple scheme provides a computational efficient extension of ICA to provide supervised classification. A more elaborate ICA mixture classifier, which is trained using a likelihood framework, is presented in [22].

EXPERIMENTS

Data

The combined image and text database is obtained from the Internet by searching for images and downloading adjacent text. The adjacent text is defined as up to 150 words one HTML paragraph tag `<P>` above or below the image, or within the row of a `<TABLE>` tag. For consistency, only jpeg were retrieved and we discarded images less than 72×72 pixels or pages without text. Three categories/classes of text/images were considered: Sport and Aviation and Paintball. Sport and Aviation categories were retrieved from www.yahoo.com (17/04/2001) and the Paintball category from www.warpig.com (21/02/2002) starting from the directories and following links until depth 5:

Category	Directory
Sports	recreation & sports → sports → pictures
Aviation	business & economy → transportation → aviation → pictures
Paintball	paintball → gallery → tournament

400 data from each category were downloaded resulting in a total of 1200 data sample, which were divided intro training and test sets of $3 \cdot 200$ samples each. Features were extracted as described above and resulted in 192 image texture features, 768 image color features, and 3591 text features (terms). In Fig. 2 examples of images from the categories are displayed.

ICA Classification

The test set classification confusion matrices obtained by using the probabilistic ICA classification scheme[4] described above are depicted in Fig. 3.

[4] The source code the deployed ML-ICA algorithm is available via the DTU:Toolbox [17].

ICA classification is done for single feature groups: texture (IT), color

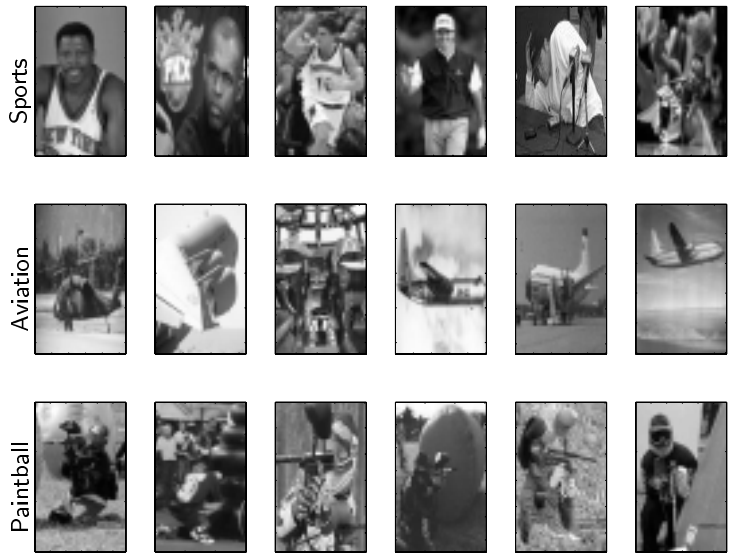
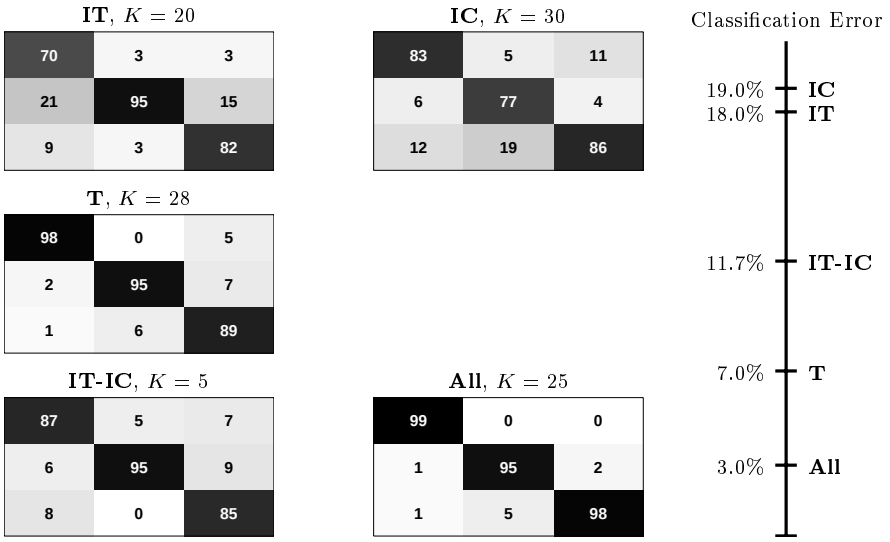


Figure 2: Images examples from the categories Sports, Aviation and Paintball.

(IC), text (T), as well 1aman7igne combinations texture-color and all features (texture-color-text). Fig. 3 (right) further shows the order of importance of the different feature groups as expressed by the overall classification error, and indicates the importance of extracting meaningful information. In this data set text features convey much more content information as compared to image features - both individually and in combination (texture-color). However, by combining all features the classification error is reduced approx. by a factor of 2 relative to using only text features. This indicates that the ICA classifier is able to exploit synergy among text and image features.

Image annotation application

An application of the suggested method is automatic annotation of text or keywords to new (test) images. In case we do not have available class labels we aim at assigning the image to a component by $\max_k p(k|\boldsymbol{x}_I)$. However, since $\boldsymbol{x}_T$ is unknown we in principle need to impute the missing value as $p(k|\boldsymbol{x}_I) = \int p(k|\boldsymbol{x}_T, \boldsymbol{x}_I) p(\boldsymbol{x}_T)\, d\boldsymbol{x}_T$. The imputation might be carried out by Monte Carlo integration, however, in this work we resort to the simple approximation $p(k|\boldsymbol{x}_I) \approx p(k|\langle \boldsymbol{x}_T \rangle, \boldsymbol{x}_I)$, where $\langle \boldsymbol{x}_T \rangle$ is the mean value of the text features on training examples. If class labels are available, we can further assign class label by $\max_y p(y|\boldsymbol{x}_I)$. In both cases associated descriptive keyword can by generated as described earlier.



Legends: **T**: text, **IT**: texture, **IC**: color, **All**: texture-color-text

Figure 3: Optimal test classification confusion matrices (left) obtained by selecting the number of components, K, to minimize the classification error (right). Repeated runs over different test sets shows that the optimal number of components has little variation. Rows and columns are estimated and correct classes, respectively, and the confusion reported in per cent sum to 100% column-wise. Rows/columns 1 through 3 correspond to Sports, Aviation and Paintball classes.

CONCLUSION

We suggested to use independent component analysis to extract meaningful content from combined text and image features, which has potential for

Image	Label	Keywords
I_1	Sports	position college weight height born lbs
I_2	Aviation	air boeing
I_3	Paintball	dogs naughty aftershock farside

Figure 4: Annotation of 3 images not used for training the model. Keywords for I_3 are team names.

web search and retrieval applications. It was demonstrated that the synergy among text and image features leads to better classification performance when using a simple probabilistic ICA classifier. The common independent component space thus convey useful information related to the content of image and adjacent text information. Finally, we provided an application example of automatic annotation of text to images using the suggested ICA framework.

Acknowledgments. This work is partly funded by the Danish Research Councils through the THOR Center for Neuroinformatics and the Center for Multimedia.

REFERENCES

[1] F. Bach and M. I. Jordan, "Kernel independent component analysis," Technical report csd-01-1166, **Computer Science Division, UC Berkeley**, 2001, http://www.cs.berkeley.edu/~jordan/papers/KernelICA.ps.gz.

[2] A. Bell and T. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," **Neural Computation**, vol. 7, pp. 1129–1159, 1995.

[3] A. Bell and T. Sejnowski, "Edges are the Independent Components of Natural Scenes," in M. C. Mozer, M. I. Jordan and T. Petsche (eds.), **Advances in Neural Information Processing Systems**, The MIT Press, 1997, vol. 9, p. 831.

[4] J.-F. Cardoso, "The three easy routes to independent component analysis; contrasts and geometry," in **Proceedings of ICA'2001**, San Diego, USA, December 9-13, 2001.

[5] M. L. Cascia, S. Sethi and S. Sclaroff, "Combining textual and visual cues for content-based image retrieval on the world wide web," in **IEEE Workshop on ContentBased Access of Image and Video Libraries**, IEEE Computer Society, 1998, pp. 24–28.

[6] S. Deerwester, S. Dumais, G. Furnas, T. Landauer and R. Harshman, "Indexing by Latent Semantic Analysis," **J. Amer. Soc. for Inf. Science**, vol. 41, pp. 391–407, 1990.

[7] D. Dunn, E. Higgins and E. William, "Optimal Gabor Filters for Texture Segmentation," **IEEE Transactions on Image Processing**, vol. 4, no. 7, pp. 947–964, 1995.

[8] J. Eakins, "Towards intelligent image retrieval," **Pattern Recognition**, vol. 35, pp. 3–14, 2002.

[9] J. Eakins and M. Graham, "Content-based Image Retrieval," Techn. report, **University of Northumbria at Newcastle**, 1999, http://www.unn.ac.uk/iidr/report.html.

[10] J. Feder, "Towards image content-based retrieval for the World-Wide Web," **Advanced Imaging**, vol. 11, no. 1, pp. 26–29, 1996.

[11] M. Flickner et al., "Query by image and video content: the QBIC system," **IEEE Computer**, vol. 28, no. 9, pp. 23–32, 1995.

[12] A. Gupta et al., "The Virage image search engine: an open framework for image management," **Storage and Retrieval for Imagea and Video**

Databases IV, vol. Proc SPIE 2670, pp. 76–87, 1996.

[13] L. K. Hansen, J. Larsen and T. Kolenda, "On Independent Component Analysis for Multimedia Signals," in L. Guan, S. Kung and J. Larsen (eds.), **Multimedia Image and Video Processing**, CRC Press, pp. 175–199, Sep. 2000.

[14] P. Højen-Sørensen, O. Winther and L. K. Hansen, "Mean Field Approaches to Independent Component Analysis," **Neural Computation**, pp. 889–918, 2002.

[15] P. Hoyer and A. Hyv, "Independent component analysis applied to feature extraction from colour and stereo images," **Computation in Neural Systems**, vol. 11, no. 3, pp. 191–210, 2000.

[16] H. Knutsson and G. Granlund, "Texture analysis using two-dimensional quadrature filters," in **IEEE Workshop CAPAIDM**, Pasadena, CA, 1983.

[17] T. Kolenda et al., "DTU:Toolbox," Internet site, **Informatics and Mathematical Modelling, Technical University of Denmark**, 2002, http://eivind.imm.dtu.dk/toolbox.

[18] T. Kolenda, L. Hansen and J. Larsen, "Signal Detection using ICA: Application to Chat Room Topic Spotting," in **Proceedings of ICA'2001**, San Diego, USA, December 9-13, 2001.

[19] T. Kolenda, L. Hansen and S. Sigurdsson, "Independent Components in Text," in M. Girolami (ed.), **Advances in Independent Component Analysis**, Springer-Verlag, pp. 229–250, 2000.

[20] H. Lappalainen, "Ensemble learning for independent component analysis," in **Proceedings of ICA'99**, Aussois, France, 1999, pp. 7–12.

[21] T.-W. Lee, M. Girolami, A. Bell and T. Sejnowski, "A Unifying Information-theoretic Framework for Independent Component Analysis," **Int. Journ. on Comp. and Math. with Appl.**, vol. 31, no. 11, pp. 1–21, March 2000.

[22] T.-W. Lee, M. Lewicki and T. Sejnowski, "Unsupervised Classification with Non-Gaussian Mixture Models using ICA," in **Proc. of Advances in Neural Information Processing Systems 11**, Cambridge MA: MIT Press, 1999.

[23] B. MacLennan, "Gabor Representations of Spatiotemporal Visual Images," Techn. Report CS-91-144, **Computer Science, Univ, of Tennessee**, 1994.

[24] C. D. Manning and H. Schütze, **Foundations of Statistical Natural Language Processing**, Cambridge, Massachusetts: MIT Press, 1999.

[25] J. Martínez, "Overview of the MPEG-7 Standard (version 5.0)," Techn. report, **ISO, Coding of moving pictures and audio**, 2001, http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm.

[26] K. Nigam, A. McCallum, S. Thrun and T. Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM," **Machine Learning**, vol. 39, pp. 103–134, 2000.

[27] Z. Pečenović, **Image retrieval using latent semantic indexing**, Master's thesis, AudioVisual Communications Lab, Ecole Polytechnique F'ed'erale de Lausanne, Switzerland, 1997.

[28] M. Pötzsch and M. Rinne, "Gabor Wavelet Transformation," Internet, 1996, http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/research/computerVision/imageProcessing/wavelets/gabor/contents.html.

[29] G. Salton, **Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer**, Addison-Wesley, 1989.

[30] T. Westerveld, "Image Retrieval: Content versus Context," in **Content-Based Multimedia Information Access, RIAO 2000 – C.I.D.-C.A.S.I.S.**, 2000, pp. 276–284.